Structural Diversity of Protein Segments Follows a Power-Law Distribution

Yoshito Sawada and Shinya Honda
National Institute of Advanced Industrial Science and Technology, Tsukuba, Japan

ABSTRACT The local structures of protein segments were classified and their distribution was analyzed to explore the structural diversity of proteins. Representative proteins were divided into short segments using a sliding *L*-residue window. Each set of local structures consisting of consecutive 1–31 amino acids was classified using a single-pass clustering method. The results demonstrate that the local structures of proteins are very unevenly distributed in the protein universe. The distribution of local structures of relatively long segments shows a power-law behavior that is formulated well by Zipf's law, implying that a protein structure possesses recursive and fractal characteristics. The degree of effective conformational freedom per residue as well as the structure entropy per residue decreases gradually with an increasing value of *L* and then converges to constant values. This suggests that the number of protein conformations resides within the range between 1.2^L and 1.5^L and that 10- to 20-residue segments are already proteinlike in terms of their structural diversity.

INTRODUCTION

Many protein structures, more than 3×10^4 in total, are now available in a public database. However, this number is negligible in comparison with the vast size of the structure space of the protein universe. For instance, a polypeptide consisting of 100 residues could theoretically have 8^{100} ($\sim 10^{90}$) different folds, assuming that the degree of conformational freedom of backbone per residue is 8 (1). The infinitesimal ratio of known structures to possible structures would be attributed to three limitations. One is an evolutional constraint. That is to say, nature has not completed the test which of all possible structures is adequate to proteins; otherwise some structures that had lived in the past became extinct accidentally. If this reason is dominant, there is a possibility that proteins having a novel fold will emerge or revive in future evolution. Another is a technical limitation. If the second reason is dominant, we will discover many new folds according to the technical improvements in structural biology. The third is a physical restriction. If the third reason is dominant, we cannot expect to find many new folds in future. This implies that the current database already contains almost all kinds of protein folds. To illuminate what regulates the number of structures, we decided to investigate the diversity of protein structures, because we hypothesize that the evolutional constraint and/or the physical restriction must influence structural diversity. We further believe that comprehensive knowledge of the diversity of protein structures will help us to develop a technology for discriminating meaningful structures, such as a protein, from mathematically possible but garbage structures.

The first taxonomic investigations into the diversity of protein structures were carried out by Murzin and colleagues, who constructed the SCOP database (2), followed by Orengo and colleagues, who developed the CATH database (3). Several second-generation classifications were generated by computational programs such as DALI (4), VAST (5), and CE (6). Recently, Kim and co-workers showed a three-dimensional map of the known folds projected onto the structure space of the protein universe (7). In all cases, the proteins were divided into domains and the domain structures were classified. Using these taxonomic classifications we can evaluate structural relationships between proteins deposited in the Protein Data Bank (PDB). Nevertheless, these classifications do not provide sufficient information to determine the genuine distribution of protein structures, because the number of available domains is too small in comparison with the structure space of the protein universe. In fact, the number of folds is not identical but dependent on the classification methods (8). Consequently, the current domain-based classification does not designate how many conformations within the vast structure space of polypeptides correspond to feasible protein structures (9).

The objectives of this study are to classify the structures of protein segments, which are smaller than domains, and to analyze their distribution. The smaller size of segments enables us to increase the sampling density, which will increase the accuracy of the statistical analysis and provide a robust distribution. Although a domain has been classically regarded as a basic unit of proteins (10), several studies have provided an alternative viewpoint describing proteins as hierarchical objects consisting of substructures smaller than domains (11–19). Accordingly, the classification of segment structures would provide insights into the molecular architecture of these hierarchical objects. Several principles for defining substructures inside a domain have been proposed so far, but none of them are universally accepted. Thus, no a priori definitions of substructures were adopted in this study.

Submitted October 26, 2005, and accepted for publication May 9, 2006.

Address reprint requests to Shinya Honda, National Institute of Advanced Industrial Science and Technology (AIST), Central 6, Tsukuba, 305-8566 Japan. E-mail: s.honda@aist.go.jp.

0006-3495/06/08/1213/11 $2.00

doi: 10.1529/biophysj.105.076661

A series of L-residue long segments was generated by sliding a clipping window sequentially along target protein sequences (L-tuple analysis). Each set of local structures consisting of consecutive L amino acids was classified by a single-pass clustering method (20), which is one of unsupervised nonhierarchical clustering algorithms. Structural dissimilarity between segments was defined on the basis of backbone dihedral angles. Two data sets containing target proteins were used. Each was a subset of the PDB from which redundant structures and low resolution data were removed. Both data sets were analyzed independently and compared to obtain reliable statistical results.

The results of these exhaustive analyses, reported here, are that the structures of protein segments are located only in tiny regions of the protein universe and distributed in a dense-sparse manner. Also, their diversity follows a power-law distribution. This indicates that proteins are organized on a certain mathematical regulation using a limited number of local structures. Moreover, our analysis of the clusters of classified segments revealed that the limitation of the number of local structures is not only due to the conformational preference of single residues. These results are an attractive outcome because they are quite similar to those found in the structure of natural languages.

METHODS

The atomic coordinates of representative proteins were obtained from the PDB, and their backbone dihedral angles were calculated. To eliminate low-resolution and/or redundant data in the PDB, we referred to two data sets. One is the PDB Select (Sep. 25, 2001, version) (21), which contains 1614 chains (resolution <3.0Å; R-factor <0.3; sequence identity <25%). The other is the Culled PDB (Dec. 13, 2001, version) (22), which contains 370 chains (resolution <1.6 Å; R-factor <0.2; sequence identity <25%). Structural dissimilarity D between two segments A and B is defined on the basis of backbone dihedral angles:

$$D=\sqrt{\frac{\sum_{l=1}^{L}[\cos^{-1}\{\cos(\phi_l^{\mathrm{A}}-\phi_l^{\mathrm{B}})\}]^2+\sum_{l=1}^{L}[\cos^{-1}\{\cos(\psi_l^{\mathrm{A}}-\psi_l^{\mathrm{B}})\}]^2+\sum_{l=1}^{L}[\cos^{-1}\{\cos(\omega_l^{\mathrm{A}}-\omega_l^{\mathrm{B}})\}]^2}{3L}}, \tag{1}$$

where L is the segment length. Cosine and arccosine functions are used to convert the difference between two angles into a value within the range 0–180°. To investigate the dependence of the segment length, the L-tuple analyses were conducted by assigning various values (1, 3, 5, 7, 8, 9, 11, 13, 15, 17, 19, 21, and 31) to L. The local structures of the segments consisting of consecutive L amino acids were classified by a single-pass clustering method (20), the flow chart of which is shown in Fig. S2 in Supplementary Materials. Briefly, 1), choose a segment and declare it to be in a cluster of size one; 2), choose a next segment and compute distances from this segment to the centroids of all clusters; 3), add the segment to the ''nearest'' cluster. If no cluster is really close (within a certain threshold), declare the segment to be in a new cluster; and 4), go to the second step unless all segments are classified. All parameters characterizing the distribution of the local structures, such as the total number of clusters N and the number of segments in the rth cluster M_r, were directly determined by this method by assigning an arbitrary value to a threshold variable for structural dissimilarity, D_{th}. In most cases, three sets of calculations were carried out after assigning 20°, 30°, or 40° to D_{th}. To obtain control distributions, pseudosegments were also classified by the same clustering method. The pseudosegments were generated by randomly selecting the values of the backbone dihedral angles calculated from the corresponding protein data sets. Each of pseudosegments has a fictitious structure, but the overall dihedral angle distribution of them is identical to that of ''real'' segments.

A normalized frequency of occurrence f_{cls} can be obtained empirically by the equation, $f_{\mathrm{cls}}(r)=M_r/M$, where M and r are the total number of segments and the rank of each cluster, respectively. The rank r corresponds to the decreasing order of f_{cls}. To hold the statistical confidence, data from the clusters whose f_{cls} were $<2.0\times10^{-5}$ were discarded, and the ''effective'' number of total clusters N_{cls} was recounted. (Note that N_{cls} is smaller than the raw number N.) As described in Results, f_{cls} shows a fine power-law distribution except for the region around $r=1$. To characterize the distribution, we also calculated an ''estimated'' frequency f_{est} by fitting f_{cls} to the Mandelbrot formula (generalized Zipf's law) with some modifications. The definition of f_{est} is as follows.

$$f_{\mathrm{est}}(r)=\begin{cases} f_{\mathrm{cls}}(r) & : r=1 \\ a(r+b)^{-\beta} & : r\geq 2 \end{cases}. \tag{2}$$

In the fitting calculation, three parameters, a, b, and β in Eq. 2, were determined numerically by minimizing Eq. 3.

$$\min_{a,b,\beta}\left[\sqrt{\frac{1}{N_{\mathrm{cls}}}\sum_{r=2}^{N\mathrm{cls}}\frac{[f_{\mathrm{cls}}(r)-f_{\mathrm{est}}(r)]^2}{f_{\mathrm{est}}(r)}}\right]. \tag{3}$$

The equation can be interpreted on the assumption that the expected error in $f(r)$ should be proportional to the square root of $f(r)$. The use of the equation is effective for obtaining a good fitting result in a double-logarithmic scale plot. Details as to the objective function in fitting calculations are described in Supplementary Materials. To compensate for the underestimation of empirical frequency of the rarest substructures, which arises from the ''zero frequency problem,'' we introduced another parameter, the estimated number of total clusters N_{est} and computed it by using the following equation,

$$N_{\mathrm{est}}=\max[n] \quad \text{subject to} \sum_{r=1}^{n} f_{\mathrm{est}}(r)\leq 1, \tag{4}$$

which was diverted from the estimating method of the number of English words by Shannon (23). Because $f_{\mathrm{est}}(r)$ is a probability density function in principle, its sum total should become unity. Structure entropy S_{est} was calculated based on N_{est} and f_{est}.

$$S_{\mathrm{est}}=-\sum_{r=1}^{N_{\mathrm{est}}} f_{\mathrm{est}}(r)\log_2 f_{\mathrm{est}}(r). \tag{5}$$

Hydrogen bonds between backbones were assigned using the DSSP (24). Only the bonds whose stabilization energy exceeded 1 kcal/mol were taken

into account. When a segment did not contain both H-donor and H-acceptor atoms, the number of hydrogen bonds was treated as 0.5. The backbone root-mean-square (RMS) deviation (Å) represents the deviation in Cartesian coordinates of three atoms (N, Cα, and C) in assigned segments from their centroidal positions (cluster center). The radius of gyration R_G (Å) is calculated with the Cα coordinates of the cluster center. To detect the amino acid preference in each cluster, the Kullback-Leibler relative entropy $\langle I(p(r)|p_0)\rangle$ is defined by the following formula (25):

$$\langle I(p(r)|p_0)\rangle = \frac{1}{L}\sum_{l=1}^{L}\sum_{E}^{20} p(r,i,l)\log_2\left(\frac{p(r,i,l)}{p_0(i)}\right), \qquad (6)$$

where $p(r,i,l)$ and $p_0(i)$ are the frequency of amino acid i at the residue position l in the rth cluster and the global frequency of amino acid i, respectively. A cluster with small $\langle I\rangle$ allows various combinations of amino acid sequences.

A pseudocluster consisting of M_r number of segments that have no structural similarity among them was also examined as a control group to analyze a statistical significance of physicochemical properties of ''real'' clusters. Pseudoclusters were generated by randomly choosing a segment M_r times from the same set of segments that was used in clustering calculation.

RESULTS

Distribution of local structures

A typical distribution of local structures is shown in Fig. 1 *a*, which was obtained by a single-pass clustering of nine-residue segments. The vertical axis shows a normalized frequency of occurrence of clusters, f_{cls}. This parameter is defined as the number of segments assigned to each cluster divided by the total number of segments in the data set. The horizontal axis represents a cluster ranking r, corresponding to the decreasing order of f_{cls}. The nearly straight lines in the double-logarithmic scale plot show conclusively that the local structures of proteins obey neither normal nor uniform distribution functions. The results also indicate that protein molecules consist of several types of common substructures showing high f_{cls} values, and numerous kinds of rare substructures showing low f_{cls} values. As to the nine-residue segments, the clusters ranked in the top 10 or 100 include 25 or 47% of the total segments, respectively, whereas the clusters ranked below 1000 contain only 22% or less. These values were not sensitive to the clustering methodologies. Before proceeding with in-depth analyses, we checked the reproducibility of clustering results and confirmed that resultant parameters are almost independent of the order of sampling in our single-pass clustering method (Figs. S3 *a* and S4, and Table S1 in Supplementary Materials). In addition, the difference between the result of the single-pass clustering and the result of an iterative calculation (100 times) based on *k*-means algorithm using the former result as an initial condition was not significant (Fig. S3 *b* in Supplementary Materials). Thus, we considered that an iterative and time-consuming calculation is not necessary for carrying out our purpose.

In Fig. 1 *a*, two distribution curves independently obtained from two PDB-derived data sets are presented together. The

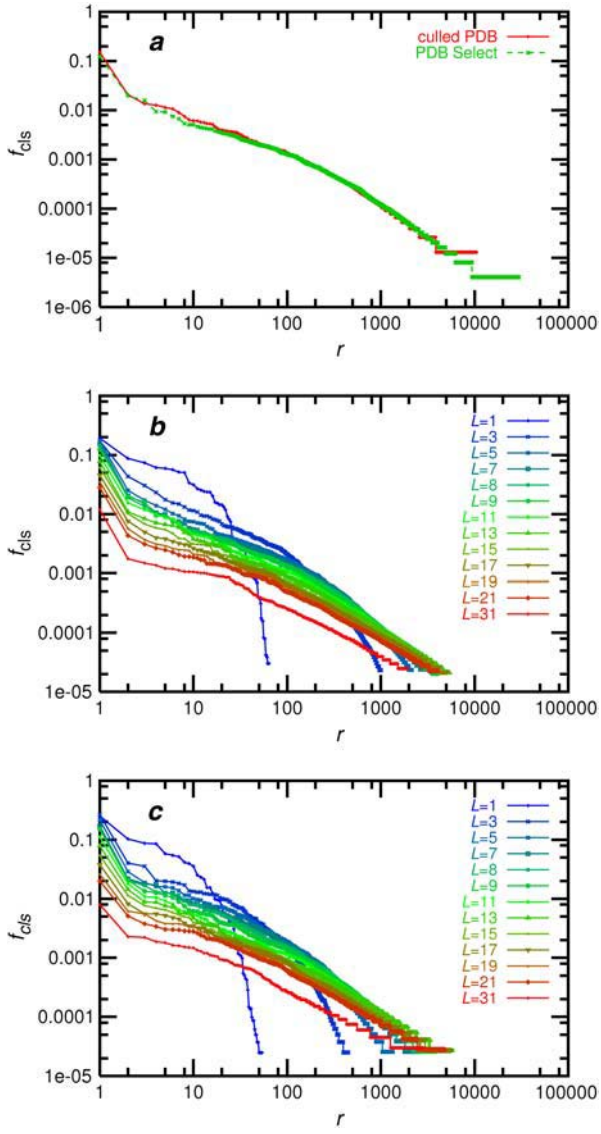

FIGURE 1 Distributions of local structures of protein segments. The normalized frequency of occurrence f_{cls} is plotted against the rank of clusters r. (*a*) Distribution curves obtained from the PDB Select (21) data set (*green*) and the Culled PDB (22) data set (*red*). Conditions: $L = 9$, $D_{th} = 30°$. (*b*) The influence of the segment length L on the distributions obtained from the PDB Select. (*c*) The distributions obtained from the Culled PDB.

difference in size between these data sets was more than a factor of 3. In fact, at $L = 9$, there were 245,894 and 76,694 segments in the large and the small data sets, respectively. Nevertheless, these two curves coincided. For instance, the f_{cls} values of the most frequent local structure ($r = 1$) obtained from the large and small data sets are 0.12 and 0.14, respectively. Thus, the agreement between these two distribution curves indicates that the number of segments (i.e., the sampling size) was statistically sufficient to obtain robust statistical results.

The only notable deviation between the two curves is seen in the lower right region of Fig. 1 *a*. This region corresponds

to the rarest substructures, which appear only once or twice among thousands of segments, so the empirical frequencies of these substructures are sensitive to the number of segments analyzed (the so-called ''zero-frequency problem'' (26)). In fact, despite considerable agreement between the two curves, the total number of clusters N (maximum value of r) differed almost threefold (Table S1 in Supplementary Materials). To eliminate this unessential complication, we excluded clusters with f_{cls} values less than 2.0×10^{-5} and introduced another parameter, the ''effective'' number of total clusters N_{cls}. This cut-off handling resulted in N_{cls} values that were comparable for the two data sets (Table S1 in Supplementary Materials).

To investigate the influence of the chain length on the distribution functions of local structures, the same type of clustering analysis was conducted after assigning various values (1–31) to L. The resultant curves are shown in Fig. 1, b and c, and the parameters are summarized in Table 1. Although some differences are seen between Fig. 1, b and c, for short segments ($L < 7$), the distribution curves for long segments ($L > 7$) are quite similar. Likewise, the relative differences in N_{cls} between the two data sets are within 15% for long segments (Table 1). Accordingly, the effect of data set size is less pronounced in the case of long segments.

TABLE 1 Influence of segment length on the number of clusters of local structure and Mandelbrot parameters of fitted curves

L	M	N_{cls}	a	b	β	N_{est}
PDB select						
1	263519	64	–	–	–	–
3	259049	1004	–	–	–	–
5	254624	2152	0.5734	18.6248	1.2331	2.5×10^3
7	250238	2919	0.3512	18.5573	1.1518	5.7×10^3
8	248061	3248	0.2313	14.9241	1.0895	8.2×10^3
9	245894	4067	0.1820	11.4119	1.0616	1.3×10^4
11	241599	4866	0.0888	6.5335	0.9621	2.5×10^4
13	237367	5385	0.0599	6.5753	0.9133	4.6×10^4
15	233188	5418	0.0393	4.7861	0.8668	7.5×10^4
17	229052	5126	0.0382	7.8167	0.8765	1.6×10^5
19	224970	4519	0.0389	12.3521	0.8928	4.0×10^5
21	220928	4059	0.0330	13.3872	0.8841	7.3×10^5
31	201352	1928	0.0152	11.4486	0.8598	2.0×10^7
Culled PDB						
1	80584	54	–	–	–	–
3	79600	450	–	–	–	–
5	78622	1341	6.9467	32.4722	1.6978	1.3×10^3
7	77654	2541	1.1152	21.8379	1.3630	4.1×10^3
8	77173	3208	0.4616	15.1095	1.2148	6.4×10^3
9	76694	3854	0.2117	10.5203	1.0878	8.4×10^3
11	75744	5157	0.0935	6.2032	0.9639	1.6×10^4
13	74806	5722	0.0597	4.6173	0.9051	2.9×10^4
15	73876	5871	0.0499	4.7903	0.8909	5.5×10^4
17	72946	5655	0.0480	7.0668	0.8978	1.2×10^5
19	72018	5187	0.0426	7.9601	0.8935	2.2×10^5
21	71099	4817	0.0349	8.7729	0.8759	3.2×10^5
31	66595	2847	0.0102	3.6882	0.7748	1.2×10^6

Refer to Methods for the meaning of symbols. Clustering conditions: $D_{th} = 30°$.

In Fig. 1, b and c, the distribution curve for a short segment changes its shape obviously with elongation of chain length, whereas the curve shape for longer segments does not vary significantly. The former shape reflects an exponential distribution, whereas the latter can be described by the power-law distribution. These shapes of longer segments are undoubtedly significant, because they are far different from the shapes of control distributions obtained from clustering pseudosegments (Fig. 2). It is well known in the field of natural-language research that the frequency of English words follows a power-law distribution (27). The phenomenon is known as Zipf's law. The original equation describing Zipf's law has been generalized by Mandelbrot (28). We have therefore tried to express the empirical curves corresponding to long segments using the modified Mandelbrot formula (Eq. 2). As shown in Fig. 2, the formulated curves superimposed well on the empirical curves except for the region around $r = 1$. This indicates that the local structure of protein segments is distributed in the structure space according to Zipf's law. One of the fitted parameters, β, which is known as the order of power law, shows a tendency to converge from 1.2–1.7 into 0.8–0.9 with elongation of the segments (Table 1).

Besides the region around $r = 1$, slight deviations of the empirical distributions from the modified Mandelbrot formula were always observed for the rarest substructures, as seen in the lower right region of Fig. 2. Since the amount of data in this region is sensitive to the number of segments analyzed, this chronic deviation may suggest that the zero-frequency problem (26) caused an underestimation of the frequency of the rarest substructures. Therefore, we calculated a new parameter, the ''estimated'' number of total clusters, N_{est}, by extrapolating the estimated frequency f_{est} rightward until the definite integral of the formulated distribution function reached 1 using Eq. 4, which is diverted from the estimating method of the number of English words by Shannon (23). Unlike N_{cls}, N_{est} increases monotonously with L (Table 1). The values of N_{est} for middle-length segments are almost comparable with N_{cls}, whereas those for longer segments are apparently larger than N_{cls}. The goodness of fit in the fitting calculation of the modified Mandelbrot formula to the empirical data was evaluated by parametric bootstrapping analyses (Table S3 in Supplementary Materials). Also, the statistical deviations of the fit coefficients that were independently obtained when the order of sampling was changed randomly are summarized in Table S1 in Supplemental Materials. These analyses imply that the results of the single-pass clustering provide parameters robust and reliable enough to characterize the structural diversity of protein segments.

Degeneracy of the structural diversity

If z denotes the degree of ''intrinsic'' conformational freedom per residue, the total number of possible backbone

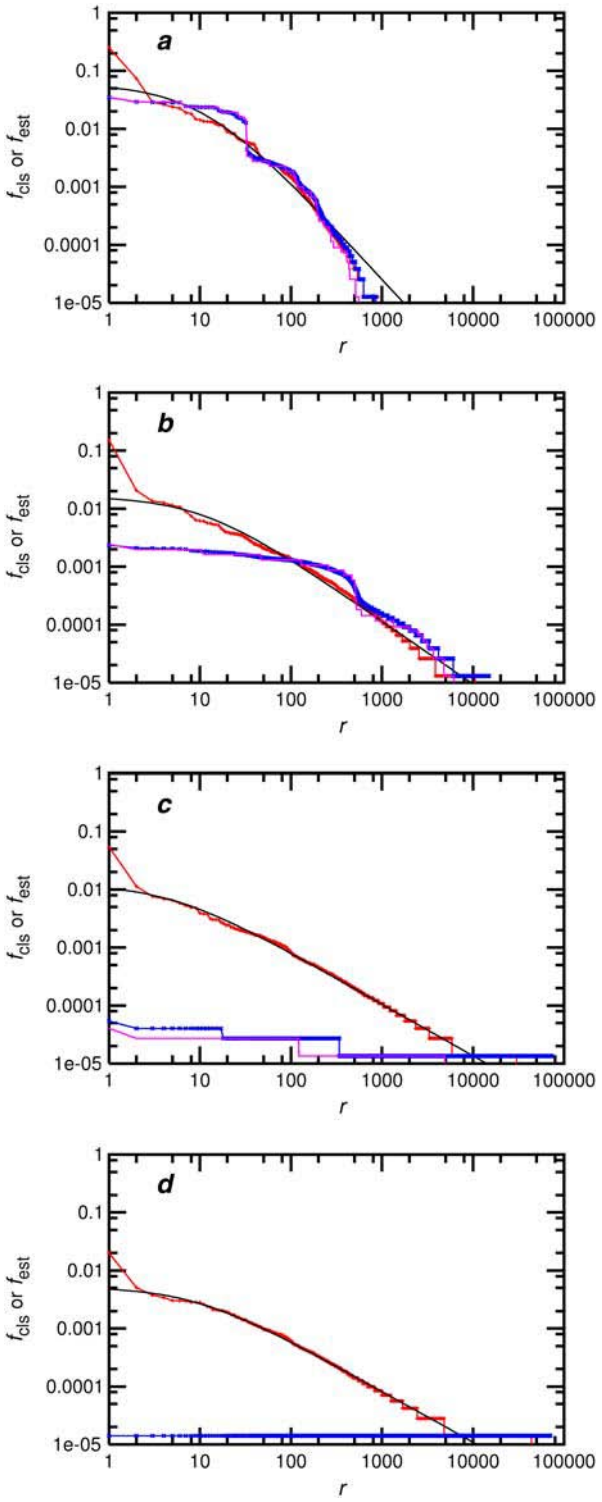

FIGURE 2 Empirical and formulated distributions of local structures. Formulated distribution curves (*black*) were obtained by the fitting calculation to the modified Mandelbrot formula. In the fitting calculations, the data at $r = 1$ of the empirical distributions (*red*) were not included, because the deviations of f_{cls} from the estimated frequency f_{est} at $r = 1$ were so large in every analysis that the fitting calculations including all data resulted in disagreement throughout the entire range (see Methods). As a control distribution, the results of the same clustering analysis of pseudosegments are also presented (*blue*). These empirical distributions of pseudosegments agree well with theoretical curves (*purple*) that were computed on the assumption of repeated permutation ($_{p}\Pi_{L}$) of four probability values (p_i) corresponding to four sets of dihedral angles; $p_1 = 0.512$, $p_2 = 0.418$, $p_3 = 0.049$, $p_4 = 0.018$. Conditions: (*a*) $L = 5$, $D_{th} = 40°$, Culled PDB; (*b*) $L = 9$, $D_{th} = 30°$, Culled PDB; (*c*) $L = 15$, $D_{th} = 30°$, Culled PDB; (*d*) $L = 21$, $D_{th} = 30°$, Culled PDB.

conformations for a fragment consisting of L residues should be z^L. If one assigns a value of 8 to z (1), the total number of conformations available to a nine-residue fragment would become 1.3×10^8. However, the estimated number of clusters for a nine-residue segment was only $\sim 10^4$ (Table 1), indicating that the actual diversity of the local structure is quite limited. Here we introduce z/α as the degree of ''effective'' conformational freedom per residue. The parameter α is a ''diminishing factor'' (29), which expresses the degeneracy of the actual diversity in the protein universe. To determine the amount of degeneracy, $(\log N_{est})/L$ were plotted against L in Fig. 3, because they are in principle equivalent to $\log(z/\alpha)$. The resultant values of z/α were not influenced appreciably by the sizes of the data sets (Fig. 3, *a* and *b*). In contrast, these values were obviously affected by L, proving that the degeneracy depends on the segment length. The plot of z/α versus L appears as an asymptotic curve. The value of z/α decreases gradually and monotonously with increasing L. The value appears to reach a constant level at $L = 31$, though the absolute value depends on a threshold parameter D_{th}: $z/\alpha =$ 1.6–1.7 and 1.5–1.6 at $D_{th} = 30°$ and 40°, respectively. This result enables us to predict the number of protein conformations. Using heuristic regression analysis to fit the data to a hyperbolic function showed that z/α extrapolates to 1.2–1.5 at $L = 100$ (the typical domain size of proteins). In addition, Fig. 3 suggests that the structural diversity of the middle length segment has been shrunk to considerable extent. According to thermodynamic analysis, the degree of backbone freedom for an unfolded protein is ~8 per residue (1), which is clearly larger than the z/α values of the middle length segment. Consequently, the data in Fig. 3 can be understood as follows: the decreasing curve represents the changes in structural degeneracy in passing from ''the polypeptide world'' ($z/\alpha \sim 8$) to ''the protein world'' ($z/\alpha \sim 1.3$). From this viewpoint, it is conceivable that the structural diversity of 10- to 20-residue segments is already approaching that of the protein world.

To further examine the degeneracy profile, we calculated structure entropy S_{est} as another parameter characterizing the distribution curves of local structures. Fig. 4 shows the dependence on L of the structure entropy per residue, S_{est}/L. The S_{est}/L curve resembles that of $(\log N_{est})/L$. With increasing L, S_{est}/L first decreases rapidly and then converges to a constant value. The difference in size of the data sets was not significant for determining the value of S_{est}/L, as seen in the analysis of $(\log N_{est})/L$. These similarities strengthen the previous idea that 10- to 20-residue segments

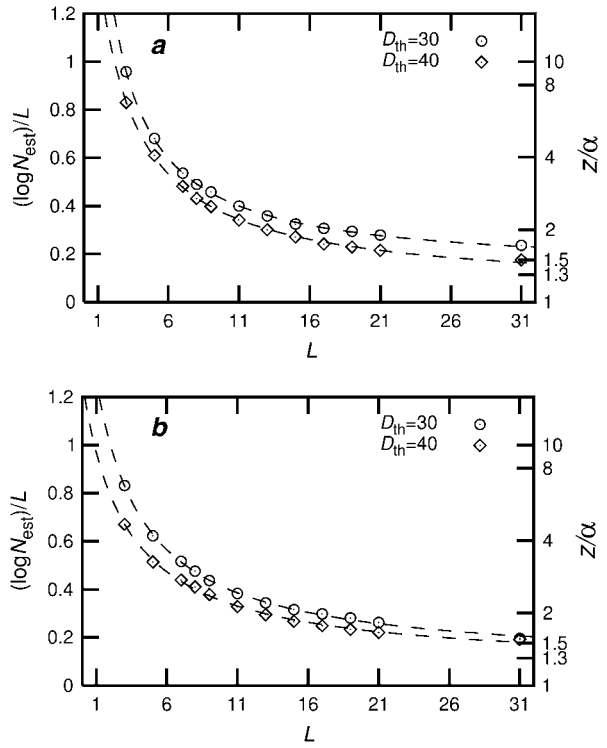

FIGURE 3 Dependence of the number of clusters per residue (logN_{est})/L on L. (Conditions: $D_{th} = 30°$ (*circles*) and 40° (*diamonds*)) The right axis indicates the degree of effective conformational freedom per residue z/α (see text). Dashed lines denote heuristically fitted hyperbolic functions. (*a*) Results from the PDB Select: $(y - 0.132)(x + 0.837) = 3.18$ for $D_{th} = 30°$ and $(y - 0.051)(x + 1.78) = 3.74$ for $D_{th} = 40°$. (*b*) Results from the Culled PDB: $(y - 0.086)(x + 2.40) = 4.02$ for $D_{th} = 30°$ and $(y - 0.074)(x + 3.03) = 3.60$ for $D_{th} = 40°$.

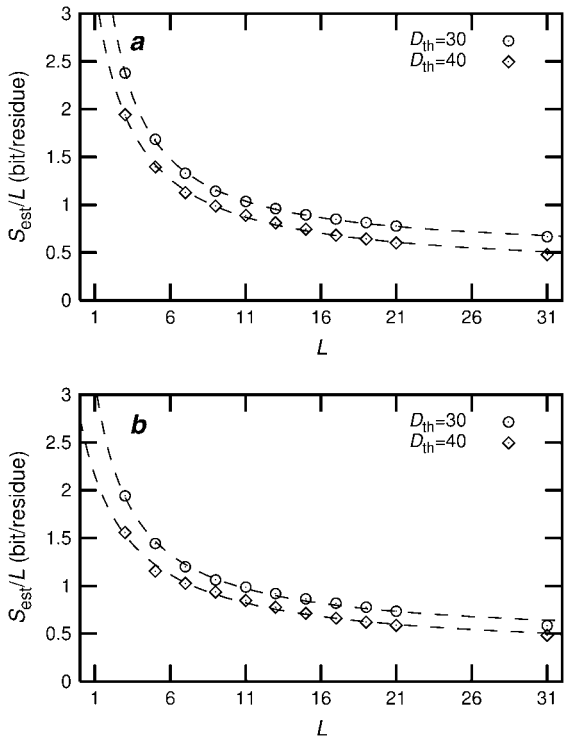

FIGURE 4 Dependence of the structure entropy per residue S_{est}/L on L. Conditions: $D_{th} = 30°$ (*circles*) and 40° (*diamonds*). (*a*) Results from PDB Select. (*b*) Results from Culled PDB.

already possess a protein-like nature in terms of structural diversity. The S_{est}/L value at $L = 31$ is 0.5–0.7 bits per residue. This suggests that one bit of information per residue would be enough to express all feasible folds/topologies of existent proteins.

Localization of local structures in the protein universe

All the analyses described above were performed with a constant threshold parameter, which is responsible for making a new cluster in our algorithm (see Methods). Therefore, it is likely that the clustering results are influenced by the value of D_{th}. We have chosen the value of 30° or 40° as an appropriate threshold parameter after a preliminary clustering analysis. The results of the preliminary analysis are shown in Fig. 5 *a*, illustrating the number of segments of the largest cluster (M_1) as a function of D_{th}. As a control, the same plots obtained from pseudosegments were also presented. In principle, the classification using a smaller threshold will lead to a decrease in the number of segments of the largest cluster, whereas a larger threshold will increase this number. In fact, in the ranges of $D_{th} = 0$–20° and 60–100°, M_1 for real segments was considerably altered by changing D_{th}. (All segments were classified into one cluster when using $D_{th} \geq 100°$.) In contrast, this number was relatively stable between 30° and 40°. This indicates that the dependence of a threshold parameter on the clustering results is minimum around $D_{th} = 30$–40°. In Fig. 5 *b* we can see that the shapes of the distribution of 13-residue segments are quite similar to each other when analyzed with $D_{th} = 30°$ and 40°.

A plateau stage like a landing of stairs from $D_{th} = 20$–60° for real segments in Fig. 5 *a* was never seen in the plots for pseudosegments. This plateau stage is one of characteristics of power-law distribution (30), suggesting the localization of local structures in the protein universe. The multidimensional space of the protein universe expands explosively with increasing segment length. Nevertheless, the real segments are localized at particular regions rather than dispersed uniformly throughout the protein universe, unlike pseudosegments. Increasing length therefore isolates each cluster from the others and makes the distance between clusters longer. As a result, the protein universe forms a space composed of a limited number of dense cores and vast sparse regions, like galaxies in the real universe.

Summary of classified structural motifs

An in-depth analysis of each cluster is beyond the scope of this article. Here we describe only a few clusters that were

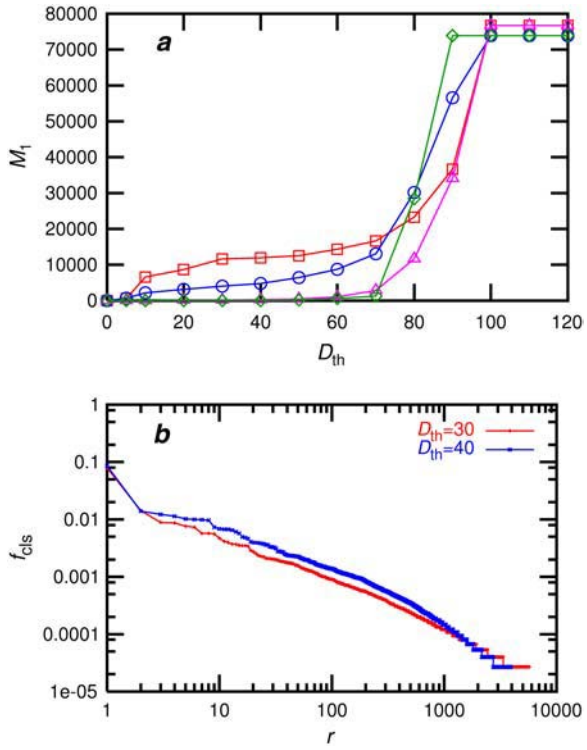

FIGURE 5 Influence of the threshold parameter D_{th} on clustering results. Conditions: Culled PDB. (*a*) Number of segments of the largest cluster M_1 as a function of D_{th}: $L = 9$ (*red*), 15 (*blue*). As a control, the same plots obtained from pseudosegments are also presented: $L = 9$ (*purple*), 15 (*green*). (*b*) Distributions of local structures of 13-residue segments: $D_{th} = 30°$ (*red*), $D_{th} = 40°$ (*blue*).

obtained under a typical condition ($L = 9, D_{th} = 30°$, Culled PDB) to illustrate that our clustering method succeeded in the extraction of distinct structural motifs, including known canonical ones (Fig. S1 in Supplementary Materials). The structure of the most frequently seen cluster at $r = 1$ is a regular α-helix having almost identical dihedral angles at every position ($\phi = -64 \pm 2.9°, \psi = -41 \pm 3.1°, \omega = 180 \pm 0.7°$). The backbone RMS deviation between assigned segments is 0.36 Å, which is the smallest deviation among the top 1000 clusters. The statistical analysis of the sequences showed that the propensity of amino acids does not exceed 1.7 for any position. This remarkable feature, i.e., minimum structural distortion with low sequence specificity, might be responsible for the large deviation of the empirical frequency f_{cls} from the estimated frequency f_{est} at $r = 1$, as shown in Fig. 2. The fully extended β-strand ($\phi = -116 \pm 6.0°, \psi = 140 \pm 5.3°, \omega = 178 \pm 1.2°$) appears in the cluster at $r = 3$. It is reasonable that overall propensities of branched hydrophobic amino acids such as Val and Ile are relatively high in this cluster. The clusters at $r = 2$ and $r = 5$ correspond to helix capping motifs. The former is the type IIb N-cap motif (31), having the consensus sequence of [Asp, Asn, Ser, Thr]-Pro-[Glu, Asp]-[Gln, Glu]. The latter is the type IV C-cap motif (31) (i.e., Schellman motif), having the consensus sequence of [Glu, Lys, Ala, Arg]-[Leu, Arg, Ala, His]-Gly. Among the many types of hairpin structures, the most frequent one appears in the cluster at $r = 79$, whose structure is the β-hairpin consisting of two β-strands and a two-residue loop. The loop corresponds to a type I′ β-turn, which is an abnormal type in the β-turn statistics. This preference for an abnormal β-turn inside the most frequently found β-hairpin has already been revealed by Sibanda and Thornton in their statistical analysis (32). Accordingly, all characteristics of the clusters presented here are coincident with previous investigations of known structural motifs, indicating that our clustering method is effective in analyzing structural motifs and is able to extract numerous motifs simultaneously and exhaustively.

Popularity of local structures

What determines ''the popularity'' of local structures, their relative occurrence in the protein universe? To address this question, we analyzed four physicochemical properties of each cluster, namely, the number of hydrogen bonds, the structural dispersion in Cartesian coordinates, the compactness, and the sequence specificity, and compared these properties to those of pseudoclusters. Fig. 6 *a* shows that the number of backbone-backbone hydrogen bonds per segment, taking account of both intra- and intersegment hydrogen bonds, correlates with the rank of clusters. This indicates that the more hydrogen bonds the local structure contains, the more frequently it occurs. In contrast, neither the number of intrasegment or intersegment hydrogen bonds shows a clear relationship with the rank (data not shown). Only the sum of hydrogen bonds has an obvious correlation with rank. These results imply that the distinction between intra- and intersegment bonds does not affect the cluster ranking, and further suggest that the individual foldability or ''autonomy'' of local structure would not be a criterion of its popularity. Next, the backbone RMS deviations of assigned segments were computed as an indication of the structural dispersion within each cluster (Fig. 6 *b*). The RMS deviations ranged from 0.4 to 2.5 Å and tended to increase with the rank. (The average, ~1.2 Å, corresponds to the level of ''good'' quality in NMR structure determinations.) This correlation can be explained by postulating that the popular local structures have little structural fluctuation and represent deep local minima on the potential energy surface. Considering the results in Fig. 6, *a* and *b*, hydrogen bonds, regardless of whether they are intra- or intersegmental, are likely to reduce fluctuations in the local structure. In contrast to the above two properties, the radius of gyration of a segment R_G, being characteristic of the compactness of the local structure, does not show a simple relationship to the rank (Fig. 6 *c*). The values of R_G for the high-ranked clusters

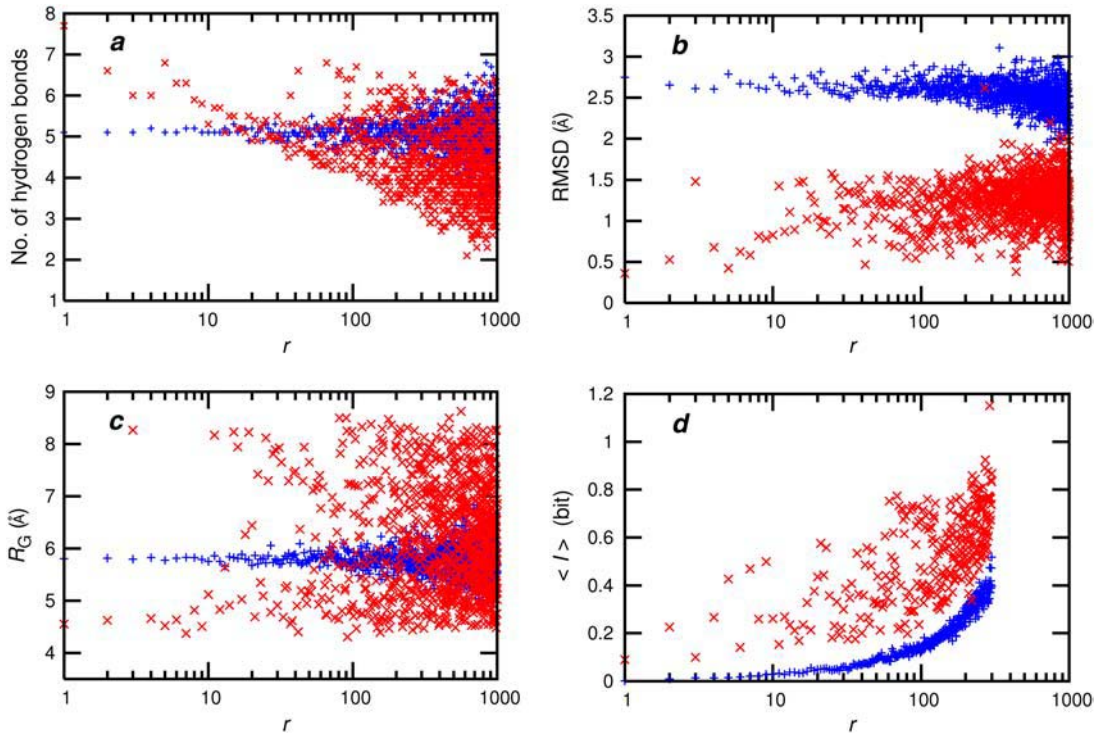

FIGURE 6 Structural and sequential properties of the top 1000 clusters (*red*). Conditions: $L = 9$, $D_{th} = 30°$, Culled PDB. Properties of pseudoclusters (*blue*) are also presented as control groups. (*a*) Average number of hydrogen bonds in segments assigned to a cluster. These values account for both intra- and intersegment hydrogen bonds. (*b*) Backbone RMS deviations of the segments from the cluster center. (*c*) Radius of gyration of the cluster center R_G. (*d*) Kullback-Leibler relative entropy of amino acid appearance $\langle I \rangle$. In view of statistical confidence, only the top 300 clusters are shown in this panel.

are displayed in a binary fashion around either 4.6 or 8.2 Å, whereas R_G for low-ranked clusters are intermediate between these two values. Because the two values correspond to the sizes of the canonical α-helix and β-strand structures, respectively, this demonstrates that neither the type of secondary structure nor the compactness of the local structure affect the popularity. In contrast, the regular secondary structures, both α-helix and β-strand, occur more frequently than complicated structures, e.g., combinations of α-helices and β-strands, as well as others. Finally, the sequence specificity of local structures was evaluated by computing the Kullback-Leibler relative entropy $\langle I \rangle$ (25), based on the propensities of position-specific amino acids. Compared with pseudoclusters, all clusters show larger values in $\langle I \rangle$, which indicates that the sequence specificity of each local structure is evidently higher than that of the same number of segments that were selected randomly (Fig. 6 *d*). In contrast, the relation between $\langle I \rangle$ and r is not simple. One of the things we can see in Fig. 6 *d* is that the values of $\langle I \rangle$ for low-ranked clusters are considerably dispersed, whereas the values for the high-ranked clusters are rather close to the level of pseudoclusters. The result would suggest that the degree of sequence specificity may also affect the popularity of local structures, and that a local structure that does not require particular amino acids tends to occur more frequently. In other words, local structures with high ''designability'' (33) may be preferred in the protein universe.

DISCUSSION

Our clustering results illustrate that the local structures of proteins are very unevenly distributed in the protein universe. In the case of long segments, the plot of f_{cls} versus r is nearly linear in the double-logarithmic scale, the shape of which is far different from that of the same plot for pseudosegments, indicating that the distribution of local structures follows a power-law distribution (Figs. 1 and 2). The power-law distribution is occasionally called the Zipf's law (27) distribution, which is also observed in the L-tuple analyses of amino acid sequences of proteins ($L = 2, 3, 4$) (34) and of DNA sequences of genes ($L = 3$–7) (35). The relationship between the nodes and links in a protein domain network also follows this type of distribution (30,36). In fields besides biology, power-law distributions are seen for various phenomena (37), such as natural languages, website popularity, Internet traffic, individual incomes, corporation sizes, and city populations. Czirok and colleagues argued that sequential binary data having a long-range correlation tend to show power-law behavior over their entire range, in contrast to short-range correlated sequences (38). Hence, the

power-law distribution of local structures can be understood by the interpretation that the combinations of dihedral angles of amino acids are not free but quite limited, and that this limitation is significant even for segments of only several residues. The limitation probably derived from long-range interactions as well as short-range interactions in a protein molecule, and it is likely to contribute to the infinitesimal ratio of existent sequences/structures against theoretical possible sequences/structures.

The structure distribution for long segments was formulated well by the modified Mandelbrot formula (Eq. 2 and Fig. 2). The order of power law was approximately unity. For instance, $\beta = 0.87$–0.89 and 0.88–0.89 at $L = 15$ and 21, respectively. These values are almost the same as the comparable parameter obtained from theoretical folding simulations using a two-dimensional HP lattice model, in which β-values were 0.94 ($L = 16$) and 0.86 ($L = 18$) (39). Interestingly, a similar convergence was reported in linguistics, in which the parameter β of the distribution of English vocabulary decreases from 1.6 to 1.15 as a child grows, and reaches 1.0 in the books of a professional novelist (40). Similar distributions appear in other natural languages, such as French and Japanese. Conventionally, the structure of a protein molecule has been analogized with a grammar (41). The results of this study imply that this resemblance is not just a metaphor and that a protein structure and a language structure probably share common rules and have a quantitative correlation.

The number of conformations of protein segments converged from 2.5^L–2.9^L at $L = 9$ to 1.5^L–1.7^L at $L = 31$ with increase of the segment length. When the data in Fig. 3 were heuristically fitted to a simple hyperbolic function, although this function has no theoretical ground to characterize the decays, the extrapolated values were between 1.2^L and 1.5^L at $L = 100$. We therefore estimate that structural diversity of proteins would reside within the range between 1.2^L and 1.7^L, depending on their chain length. Interestingly, Dill has reported similar values, 1.4^L or 1.7^L, as the upper limit of the number to conformations of globular states by using a three-dimensional lattice model (29). Recently, Kim and co-workers have also obtained an equivalent value, 1.6^L, from the analysis of short segments ($L = 2$–7) by the multidimensional scaling algorithm (42). It should be noted that these estimated numbers are comparable despite the difference in analytical method, and that all numbers are smaller than 2^L. This indicates that the number of protein structures is evidently smaller than the number of random combinations of two sets of dihedral angles corresponding to two typical secondary structures, i.e., α-helix and β-strand. On the other hand, the number of conformations of an unfolded protein has been estimated (1) to be $\sim 8^L$. Since it is reasonable to regard the structural diversity of an unfolded protein as equivalent to that of a random polypeptide, the structural degeneracy of a protein, defined as the ratio of the structure space of existent proteins against the vast protein universe, can be estimated to be $(1/7)^L$–$(1/5)^L$.

The structural diversity appears to exhibit a boundary at 10–20 residues. With increasing L, both the degree of conformational freedom per residue z/α and the structure entropy per residue S_{est}/L decrease gradually and then reach an almost constant level at $L = 10$–20 (Figs. 3 and 4), suggesting that 10- to 20-residue segments are already proteinlike and that their nature differs considerably from that of shorter segments. This can be related to several other types of research in protein science. In a fragment assembly method (43), which is one of the most accurate methods of structural prediction (44), the length of nine-residue segments is empirically known to be suitable to produce excellent results in predictions. Also, in the fields of molecular evolution and of protein folding, a protein molecule is often considered as a hierarchical object composed of smaller units than the typical size of a domain (11–19). Our results may help to explain the basis for the minimum chain length of local structural units of proteins.

As for longer segments, a double logarithmic plot of the number of clusters versus the cluster size determined by the number of assigned segments appears linear (Fig. 7), showing that this relationship also follows a power law. The order of power law γ was 2.2 for 21-residue segments. Similar relationships have been discovered recently between the numbers of folds and families (45), between the numbers of families and domains (45), and between the numbers of clusters of similar domain and domains per cluster (30). In their analyses, γ were reported as 2.5, 3.0 (or 1.9), and 2.5, respectively. These examples of power-law behavior between upper and lower categories indicate that there are recursive relationships in a protein molecule in the progression of fold-family-domain-segment-sequence. Furthermore, this hierarchical self-similarity suggests that a protein holds fractal characteristics in its structure. The fractal characteristics of protein structure have been suggested by early investigations, such as the temperature dependence of ESR spectra (46) and the differential-geometric analysis (47). As

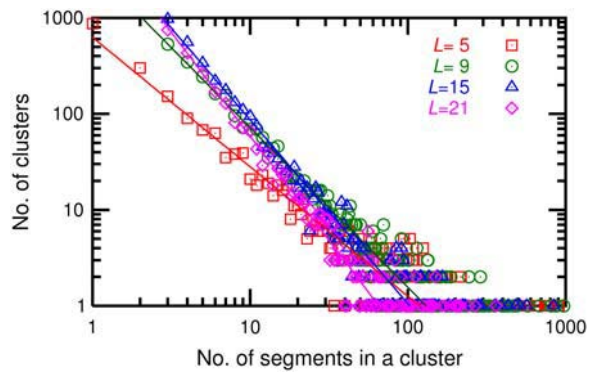

FIGURE 7 Double logarithmic plot of the number of clusters versus the number of segments in a cluster M_r. Conditions: $D_{th} = 30°$, Culled PDB. The results of the representative segment length are displayed. $L = 5$ (*red*), 9 (*green*), 15 (*blue*), and 21 (*magenta*). Solid lines denote regression results on the equation $y = ax^{-\gamma}$; $y = 627x^{-1.3}$ for $L = 5$; $y = 3316x^{-1.7}$ for $L = 9$; $y = 7061x^{-1.9}$ for $L = 15$; and $y = 8588x^{-2.2}$ for $L = 21$.

to the reason for the fractal characteristics, Allen et al. proposed that it is associated with the requirement that these linear (one-dimensional) polymers having no branches should fold into a three-dimensional compact structure (46). The fractal characteristics may underlie not only the hierarchical organization of a protein molecule but also the diversity of protein structures.

The popularity of local structure is decided by at least two factors (Fig. 6). One of these is hydrogen bonds between backbones. Linear correlation of the number of hydrogen bonds to $\log(r)$ indicates that a strongly stabilized local structure is found frequently. The correlations among the hydrogen-bond density, the RMS deviation and the popularity can be interpreted by the idea that the local structure having a deep and narrow local minimum on the free-energy surface is favored, because these factors can be considered to correspond to the terms of internal energy and conformational entropy, respectively. In addition to structural factors, sequential factor ''designability'' may also affect the popularity of local structure. A local structure with sequence that allows various combinations of amino acids appears to be preferred. The finding that diversity of the amino acid sequence is also responsible for determination of the distribution of segment structure would be important to an understanding of protein evolution, as suggested earlier by Li et al. using a simple lattice model (33).

The dependence of M_1 on D_{th} showed a plateau stage like a landing of stairs, which is not found in the same analysis for pseudosegments (Fig. 5 *a*). This implies that the distribution of real segments is not random graphlike, and that local structures are distributed in a dense-sparse manner in the protein universe. Thus, the distribution of existent proteins in the protein universe should resemble galaxies in the real universe. This galaxy model allows interpretation of protein structures by means of the simple combination of known local structures. The sparse regions between galaxies correspond to inappropriate structures for a protein. Consequently, by eliminating the sparse regions we can reduce effectively the size of candidates that must be considered without losing the actual diversity of a protein.

Two data sets were analyzed independently to validate statistical reliability in this study. The resulting distribution of local structures hardly changed between the two data sets. This fact demonstrates that representative proteins listed in two data sets were fairly selected from the PDB without bias, and suggests that these data sets cover most of all local structures that can exist in a protein molecule. Accordingly, though it is expected that many proteins having a novel fold will be discovered by structural genomics projects, we guess that their structures can be mostly expressed by combinations of the local structures clarified here. Moreover, we believe that the distribution of local structures provided in this study does not differ significantly from the genuine distribution of the local structures of all natural proteins, including ones whose structures are still unknown.

Recently, Kim and co-workers have shown a three-dimensional map of the protein-fold space that helps us to understand a global feature of the protein structure universe (7). In this study, we analyzed the diversity of proteins and their distributions by focusing on the local structures of segments ($L = 1$–31). Consequently, we conclude that the local structures of proteins are distributed according to a power law and are localized in the protein universe. Very recently, Higo and co-workers reported on conformational distribution of short segments through a different approach, a principal-component analysis using intrasegment C_α-C_α atomic distances, and discussed several structural motifs, including novel ones (48). Also, Kim and co-workers carried out a multidimensional scaling analysis of short segments ($L = 2$–7) using two dihedral angles, Φ and Ψ, and concluded a dramatic reduction of conformational space by projecting an intrinsically multidimensional space of the protein universe on a three-dimensional map (42). Their conclusion is conceptually consistent with the galaxy model of this study. We think the next question is how the distribution of local structures quantitatively correlates to the distribution of local sequences. As one of the applications, we recently succeeded in designing a small folded peptide consisting of only 10 amino acids according to the original strategy that was developed based on knowledge about the uneven distribution of local structures (49), which has been fully described in this study. We believe that a deep understanding of the correlation between the diversities of structure and sequence will encourage the advance of future studies on structure prediction and molecular evolution as well as protein design.

SUPPLEMENTARY MATERIAL

An online supplement to this article can be found by visiting BJ Online at http://www.biophysj.org.

REFERENCES

1. Privalov, P. L. 1979. Stability of proteins: small globular proteins. *Adv. Protein Chem.* 33:167–241.
2. Murzin, A. G., S. E. Brenner, T. Hubbard, and C. Chothia. 1995. SCOP: a structural classification of proteins database for the investigation of sequences and structures. *J. Mol. Biol.* 247:536–540.
3. Orengo, C. A., A. D. Michie, S. Jones, D. T. Jones, M. B. Swindells, and J. M. Thornton. 1997. CATH: a hierarchic classification of protein domain structures. *Structure.* 5:1093–1108.
4. Holm, L., and C. Sander. 1993. Protein structure comparison by alignment of distance matrices. *J. Mol. Biol.* 233:123–138.
5. Madej, T., J. F. Gibrat, and S. H. Bryant. 1995. Threading a database of protein cores. *Proteins.* 23:356–369.
6. Shindyalov, I. N., and P. E. Bourne. 1998. Protein structure alignment by incremental combinatorial extension (CE) of the optimal path. *Protein Eng.* 11:739–747.
7. Hou, J., G. E. Sims, C. Zhang, and S. H. Kim. 2003. A global representation of the protein fold space. *Proc. Natl. Acad. Sci. USA.* 100:2386–2390.

8. Liu, X., K. Fan, and W. Wang. 2004. The number of protein folds and their distribution over families in nature. *Proteins*. 54:491–499.
9. Taylor, W. R. 2002. A 'periodic table' for protein structures. *Nature*. 416:657–660.
10. Jaenicke, R. 1999. Stability and folding of domain proteins. *Prog. Biophys. Mol. Biol.* 71:155–241.
11. Gilbert, W. 1978. Why genes in pieces? *Nature*. 271:501.
12. Blake, C. C. 1978. Do genes-in-pieces imply proteins-in-pieces? *Nature*. 273:267.
13. Go, M. 1981. Correlation of DNA exonic regions with protein structural units in haemoglobin. *Nature*. 291:90–92.
14. Seidel, H. M., D. L. Pompliano, and J. R. Knowles. 1992. Exons as microgenes? *Science*. 257:1489–1490.
15. Karplus, M., and D. L. Weaver. 1976. Protein-folding dynamics. *Nature*. 260:404–406.
16. Baldwin, R. L., and G. D. Rose. 1999. Is protein folding hierarchic? I. Local structure and peptide folding. *Trends Biochem. Sci.* 24: 26–33.
17. Iwakura, M., T. Nakamura, C. Yamane, and K. Maki. 2000. Systematic circular permutation of an entire protein reveals essential folding elements. *Nat. Struct. Biol.* 7:580–585.
18. Lupas, A. N., C. P. Ponting, and R. B. Russell. 2001. On the evolution of protein folds: are similar motifs in different protein folds the result of convergence, insertion, or relics of an ancient peptide world? *J. Struct. Biol.* 134:191–203.
19. Rost, B. 2002. Did evolution leap to create the protein universe? *Curr. Opin. Struct. Biol.* 12:409–416.
20. Richards, J. A., and X. Jia. 1999. Remote sensing digital image analysis. Springer-Verlag, New York.
21. Hobohm, U., M. Scharf, R. Schneider, and C. Sander. 1992. Selection of representative protein data sets. *Protein Sci.* 1:409–417.
22. Wang, G., and R. L. Dunbrack, Jr. 2003. PISCES: a protein sequence culling server. *Bioinformatics*. 19:1589–1591.
23. Shannon, C. E. 1951. Prediction and entropy of printed English. *Bell Syst. Tech. J.* 30:51–64.
24. Kabsch, W., and C. Sander. 1983. Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers*. 22:2577–2637.
25. Kullback, S., and R. A. Leibler. 1951. On information and sufficiency. *Ann. Math. Stat.* 22:79–86.
26. Witten, I. H., and T. C. Bell. 1991. The zero-frequency problem: estimating the probabilities of novel events in adaptive text compression. *IEEE Trans. Inf. Theory*. 37:1085–1094.
27. Zipf, G. K. 1949. Human behavior and the principle of least effort. Addison-Wesley, Cambridge, MA.
28. Mandelbrot, B. 1953. An information theory of the statistical structure of language. *In* Communication Theory. W. Jackson, editor. Academic Press, New York. 486–502.
29. Dill, K. A. 1985. Theory for the folding and stability of globular proteins. *Biochemistry*. 24:1501–1509.
30. Dokholyan, N. V., B. Shakhnovich, and E. I. Shakhnovich. 2002. Expanding protein universe and its origin from the biological Big Bang. *Proc. Natl. Acad. Sci. USA*. 99:14132–14136.
31. Aurora, R., and G. D. Rose. 1998. Helix capping. *Protein Sci.* 7:21–38.
32. Sibanda, B. L., and J. M. Thornton. 1985. Beta-hairpin families in globular proteins. *Nature*. 316:170–174.
33. Li, H., R. Helling, C. Tang, and N. Wingreen. 1996. Emergence of preferred structures in a simple model of protein folding. *Science*. 273:666–669.
34. Strait, B. J., and T. G. Dewey. 1996. The Shannon information entropy of protein sequences. *Biophys. J.* 71:148–155.
35. Luscombe, N. M., J. Qian, Z. Zhang, T. Johnson, and M. Gerstein. 2002. The dominance of the population by a selected few: power-law behaviour applies to a wide variety of genomic properties. *Genome Biol.* 3:R00401–R00407.
36. Wuchty, S. 2001. Scale-free behavior in protein domain networks. *Mol. Biol. Evol.* 18:1694–1702.
37. Barabasi, A.-L. 2002. Linked: the new science of networks. Perseus Books, Cambridge, MA.
38. Czirok, A., R. N. Mantegna, S. Havlin, and H. E. Stanley. 1995. Correlations in binary sequences and a generalized Zipf analysis. *Phys. Rev. E*. 52:446–452.
39. Bornberg-Bauer, E. 1997. How are model protein structures distributed in sequence space? *Biophys. J.* 73:2393–2403.
40. Pierce, J. R. 1980. An introduction to information theory: symbols, signals and noise. Dover, New York.
41. 2002. Folding as grammar. Nat. Struct. Biol. 9:713.
42. Sims, G. E., I. G. Choi, and S. H. Kim. 2005. Protein conformational space in higher order ϕ-ψ maps. *Proc. Natl. Acad. Sci. USA*. 102: 618–621.
43. Simons, K. T., C. Kooperberg, E. Huang, and D. Baker. 1997. Assembly of protein tertiary structures from fragments with similar local sequences using simulated annealing and Bayesian scoring functions. *J. Mol. Biol.* 268:209–225.
44. Kinch, L. N., J. O. Wrabl, S. S. Krishna, I. Majumdar, R. I. Sadreyev, Y. Qi, J. Pei, H. Cheng, and N. V. Grishin. 2003. CASP5 assessment of fold recognition target predictions. *Proteins*. 53(Suppl. 6):395–409.
45. Koonin, E. V., Y. I. Wolf, and G. P. Karev. 2002. The structure of the protein universe and genome evolution. *Nature*. 420:218–223.
46. Allen, J. P., J. T. Colvin, D. G. Stinson, C. P. Flynn, and H. J. Stapleton. 1982. Protein comformation from electron spin relaxation data. *Biophys. J.* 38:299–310.
47. Isogai, Y., and T. Itoh. 1984. Fractal analysis of tertiary structure of protein molecules. *J. Phys. Soc. Japan*. 53:2162–2171.
48. Ikeda, K., K. Tomii, T. Yokomizo, D. Mitomo, K. Maruyama, S. Suzuki, and J. Higo. 2005. Visualization of conformational distribution of short to medium size segments in globular proteins and identification of local structural motifs. *Protein Sci.* 14:1253–1265.
49. Honda, S., K. Yamasaki, Y. Sawada, and H. Morii. 2004. 10 residue folded peptide designed by segment statistics. *Structure*. 12:1507–1518.